DRAFT: Gut Reaction, the Health Data Research UK data hub for Inflammatory Bowel Disease

Authors

Yuchun Ding[1], Aneeq Rehman[1], Alvaro Ullrich[1], Neil Walker (neil.walker@bioresource.nihr.ac.uk) [1,2]

Affiliations

1. Health Data Research UK, University of Cambridge
2. National Institute for Health Research BioResource, University of Cambridge

Abstract

Inflammatory Bowel Disease (IBD) affects around 500,000 people in the UK. The recent advent of artificial intelligence (AI) has demonstrated promising potential to improve diagnostic efficiency, and research reliability on varied clinical tasks. However, the large volume and complexity of healthcare data makes it difficult for data to be collected, processed, and analysed by traditional approaches. As a solution, a multidisciplinary team has been formed to deliver a powerful new platform to accelerate Crohn's and Colitis research - Gut Reaction. Funded as a Health Data Research UK (HDRUK) Hub, Gut Reaction aims to build on the high-quality health data on 35,000 participants in the NIHR IBD BioResource by combining it with 'real-world' data from participating NHS hospitals, audit and Patient Reported Outcomes Measures (PROMs) from the IBD Registry and genomic data from the UK IBD Genetics Consortium. Through Gut Reaction's partner organisations AIMES - who provide secure infrastructure - and Privitar - who provide Privacy Enhancing Technologies - researchers can apply to access robustly de-personalised data in a trusted location.

Background & Summary

Inflammatory Bowel Disease (IBD) includes Crohn's disease and ulcerative colitis. Together, these conditions affect around 500,000 people in the UK, causing recurring abdominal symptoms, which need long-term treatment and often major surgery. As a result, IBD can significantly affect the lives of those who live with it.

In recent years, the advent of artificial intelligence (AI) has demonstrated promising potential to improve diagnostic efficiency (ref), and research reliability on varied clinical tasks, such as detecting early cancer lesions, or predicting the clinical outcomes of medications. However, the inconvenient truth is that the healthcare system consists of large volumes of data which are usually generated from diverse sources such as case reports, hospital admissions and discharge summaries, medical imaging, lab results, genomics and many more. The large volume as well as the complexity of these data makes it difficult for the data to be collected, processed, and analysed by traditional approaches to fit the research needs efficiently, due to both technical and governance reasons.

As a potential solution, an alliance of clinicians, academics, research nurses, funders, coordinators, programmers and, most importantly, patients have come together in the UK to deliver a powerful new platform to accelerate Crohn's and colitis research— Gut Reaction. Gut Reaction is funded as one of Health Data Research UK's (HDRUK) Hubs (ref). It allows researchers to bring together data from three

existing well-used resources - the National Institute for Health Research (NIHR) IBD BioResource (ref), IBD Registry (ref) and UK IBD Genetics Consortium at the Wellcome Sanger Institute (ref). Tens of thousands of participants have consented to join each of these resources, with substantial overlap, and data on around 8,000 IBD BioResource participants is being supplemented with "real-world" longitudinal data from participating NHS (National Health Service) hospitals.

Datasets in Gut Reaction are complimentary, reflecting the primary purposes of the contributing organisations. The NIHR IBD BioResource brings self-report demographics and health and lifestyle data, and a clinical case report form summarising medical history. Because participants in the BioResource may be invited to take part in further studies (ref), there is an additional emphasis on managing participants' current contact details and consent options, and further data and samples are held to allow the targeting of follow-up experimental research studies. The IBD Registry has largely engaged in clinical audit and safety evaluations (ref), but here offers Patient Reported Outcome Measures (PROMs). The Wellcome Sanger Institute, through its IBD Genetics Consortium, has exome- or whole-genome- sequenced over 20,000 participants, looking for rarer genetic associations, and for predictors of disease trajectory (ref).

This **DRAFT** paper describes the NIHR IBD BioResource datasets. **Figure 1** shows the intersection of these datasets. The paper will be submitted for peer-review when the datasets from the other partners have been similarly described – we understand that the interest in Gut Reaction comes through the intersection of datasets, especially those hard-to-reach data sources in the NHS.

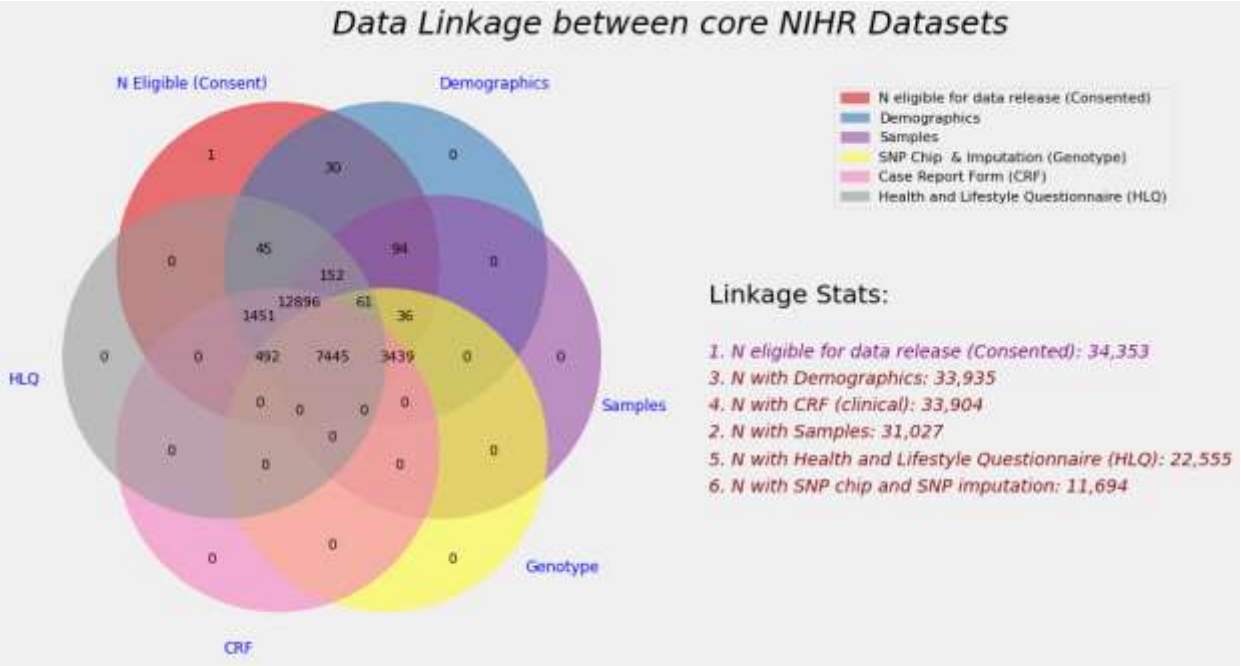


Figure 1: the intersection of NIHR IBD BioResource datasets within Gut Reaction.

By mid-2022, Gut Reaction will have created the world's largest virtual repository of data from people with IBD. Researchers can already search metadata about available datasets both at the HDR UK Innovation Gateway (ref) and on the Gut Reaction website (ref). Researchers can (ref) and do (ref) apply for access to data to support their research. Gut Reaction follows the "5 Safes" principle of data access, proposed in 2017 by the Office of National Statistics (ONS) (ref) and the UK Data Service (UKDS) (ref). It

only approves access to data by: safe people working on safe projects with safe data at a safe setting and with safe outputs. Gut Reaction partners with AIMES Management Services Ltd (AIMES)(ref) to provide secure infrastructure, and Privitar (ref) to provide Privacy Enhancing Technologies, in order that researchers can apply to access robustly de-personalised data in a trusted location.

Methods:

This section describes the various methods of data collection, storage and management. It also describes how users can apply for data access.

Data sources

The Gut Reaction Hub is collating and intersecting existing data from three different groups of consented participants. These are:

- IBD BioResource
- IBD Registry
- UK IBD Genetics Consortium

Additional linkages are sought, with NHS Trusts as well as e.g. NHS Digital, for a proportion of the IBD BioResource participants.

Data sources are summarized in **Table 1**.

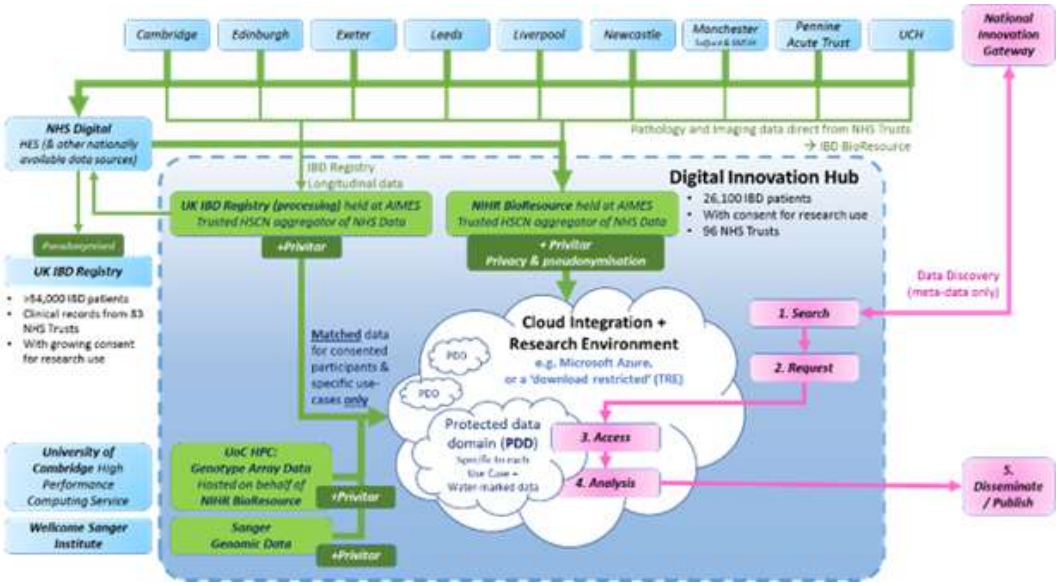
Source Institution	Programme	N Participants eligible for data release	Data Types	Additional Funders
NIHR BioResource	IBD BioResource	~34,000	1. Health and Lifestyle Questionnaires (HLQ) 2. Clinical Case Report Form (CRF) 3. Sample Holdings (DNA, Serum, Plasma) 4. Genetics (SNP Chip and SNP Imputation)	NIHR, MRC
NHS Trusts	-	~8000	1. Electronic Health Records (EHR) 2. Diagnostic (Lab Results, Imaging) 3. Prescription 4. Clinical Notes 5. Admission, Discharge Summaries	NHS England
IBD Registry	Patient Reported Outcome Measures (PROMS)- Covid-19	~58,000	Self- Reported Risks and Outcomes.	Crohn's & Colitis UK
Wellcome Sanger Institute	UK IBD Genetics Consortium	~20,000	Whole Genome/Exome Sequencing	Wellcome

Table 1: Data sources in Gut Reaction

Data collection / generation

Introduction

The Gut Reaction project operates as a secure data pipeline that covers the whole data lifecycle process. The pipeline commences with the **acquisition** from the multiple project data sources, followed by **transforming** and **loading** those datasets into one of the project audited databases, and finishes with **releasing** *ready-to-use* clean and de-personalised datasets to approved researchers. The canonical analysis space is a Trustworthy Research Environment at AIMES, with data de-personalised using Privitar's privacy enhancing technologies. A schematic for the process is shown in **Figure 2**.



**

Figure 2: schematic used to illustrate the Gut Reaction virtual repository concept to potential data partners

Methodologies for data collection / generation

Data collection / generation is ongoing during the period of this grant funding.

Data, therefore, continues to be received - at different frequencies - from the source institutions outlined above.

Data provenance is described in **Table 2**. This uses ontologies from the HDR UK Innovation Gateway - https://www.healthdatagateway.org/ - where metadata concerning these datasets is lodged. A wider metadata dataset conforms to standards used by the UK Data Archive - Metadata concerning these datasets is lodged at the HDR UK Innovation Gateway - https://www.healthdatagateway.org/. A wider metadata dataset conforms to standards used by the UK Data Archive - https://www.dataarchive.ac.uk/managing-data/standards-and-procedures/metadata-standards/ .

Survey data from NIHR BioResource and IBD Registry are taken at particular timepoints, especially recruitment, and are managed and curated by their respective data management teams. Genetic data from both NIHR BioResource and the Wellcome Sanger Institute are taken through standard QC - the former based on UK Biobank's pipeline as described in https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/genotyping_qc.pdf , the latter as described in https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7035382/

All data is received via secure routes and access occurs within a secure data centre run by AIMES Management Services Ltd (AIMES) (https://aimes.uk/). The provenance of data is listed in **Table 2**.

Source Institution	Programme	Data types	Source Data	Collection Situation
NIHR BioResource	IBD BioResource	Health & Lifestyle Questionnaire (H&LQ) Clinical Report (CRF) Demographics	Paper based, Electronic survey	Clinic, Community, Home
NIHR BioResource	IBD BioResource	Genetics (SNP, SNP Imputation) Samples (DNA, PLASMA, SERUM)	Machine generated, Laboratory Information Management System (LIMS)	Other, Clinic
NHS Trusts	*Collated by* Gut Reaction	Diagnostics Prescriptions Clinical Notes	Electronic Health Records (EHR)	Accident and Emergencies, Hospital databases, Outpatients, Inpatients
IBD Registry	COVID-19, Patient Reported Outcome Measures (PROMS)	Self-Reported Risks & Outcomes	Electronic survey	Home
Wellcome Sanger Institute	UK IBD Genetics Consortium	Whole genome/exome sequencing	Machine generated	Other

Table 2: data provenance in Gut Reaction

Data quality and standards

We adhere to the following principles of data quality:

- Accuracy – data should be sufficiently accurate for their intended purposes.
- Validity – data should be recorded and used in compliance with relevant requirements, including the correct application of any rules or definitions.
- Reliability – data should reflect stable and consistent data collection processes across collection points and over time.
- Timeliness - data should be captured as quickly as possible after the event or activity and be available for the intended use quickly and frequently enough to support information needs and to influence service or management decisions.
- Relevance – data should be relevant to the purposes for which they are used. This entails periodic review of requirements to reflect changing needs.
- Completeness – Data requirements should be clearly specified based on the information needs of the organization and data collection processes matched to these requirements.

The following therefore has been considered across all service areas:

- Staff are made aware by their line manager of their responsibilities in relation to data quality
- Commitment to data quality is clearly stated in job descriptions for all relevant roles
- Staff have the relevant skills and competencies to fulfil their role in ensuring good quality data
- Staff receive appropriate training and guidance
- Training needs are identified through the appraisal process and built into personal development
- Data quality is a key part of the induction process
- Commitment to data quality is clearly communicated through the organization.

This **DRAFT** is focussing on data from the IBD BioResource programme, which has the following systems and processes.

All clinical and administrative records must be input into approved systems. The use of any IT system to record service user data, other than those listed in **Table 3**, is to be avoided.

Name	Programme	Purpose	Dataset	Scale	Format
RedCap	IBD BioResource	Online survey tool	CRF H&LQ	GB	Relational Database Available as csv
OpenClinica	IBD BioResource	Online clinical trial management tool - *see note below*	CRF H&LQ	GB	Relational Database Available as csv
CiviCRM	IBD BioResource	Recruitment database	Demographics Consent H&LQ subset	GB	Relational Database Available as csv
Microsoft 365	IBD BioResource	Document store	Consent forms	TB	Scanned images, PDFs Not available
i2b2	IBD BioResource	Cohort discovery tool - *snapshot collation of above*	Demographics CRF H&LQ	GB	Relational Database Available as csv
University of Cambridge High Performance Computing Service	IBD BioResource	Big data computing environment	Genetic whole genome, whole exome sequence data	TB	BAMs, CRAMs & VCFs Accessed *in situ*, or via European Genome-Phenome Archive (EGA) managed access repository
proprietary LIMS	IBD BioResource	Samples database	Sample details	GB	Relational Database Available as csv

Table 3: applications and datastores used by the NIHR BioResource for data transferred to Gut Reaction. All data and samples are captured at the time of recruitment, excepting genetic data, which is generated as sufficiently large batches are assembled. All but the HPC and i2b2 (which is re-built each week as a snapshot) have audit capabilities to allow long-term curation of data. All can be output in non-proprietary formats. In creating a snapshot, i2b2 codes items to clinical ontologies: SNOMED-CT and Human Phenotype Ontology (HPO). [Table is snapshot from published DMP]

The data entry systems will be configured, where possible, to ensure that the business processes are followed.

In particular, that the system is configured to follow the participant pathway. The collection and input 'trigger points' will be identified and referenced in training materials. All changes to the clinical and administrative systems will be quality controlled to assure standards concerning the accuracy of recording data.

Fields will be made mandatory where a data item must be collected in all circumstances. The need to make further fields mandatory is kept under review subject to the necessary criteria.

Data Quality is also achieved on the bases on data types:

1. For demographic and sensitive personal information, all administration and clinical staff are responsible for checking details with the participants and volunteers at all appropriate attendances. Where changes are identified they should follow the NIHR BioResource procedures for ensuring that the change is recorded appropriately. It is vital that all demographic data is recorded accurately, completely and kept as up-to-date as possible.
2. Clinical coding is practiced in all datasets to make sure the information is of the highest standard.
3. The responsibility and ownership of data rests with the system user who must ensure that any errors are corrected promptly at source. Where validation reports are available from systems for use by clinical, managerial and data quality staff, these should be used to check for inaccurate, incomplete or untimely data.

Data Quality incidents are also part of the NIHR BioResource data quality management process. When serious data quality incidents occur or are identified, they should be reported immediately using the organizations incident reporting system and corrective action commenced.

No level of inaccuracy should be viewed as acceptable. Data quality reports are available to help staff identify data quality issues.

Careful monitoring and error correction supports good data quality. However it is more effective and efficient for data to be entered correctly in the first instance. In order to help achieve this, procedures must exist within the BioResource so that staff can be trained and supported in their work.

Situations that could arise due to insufficient information being recorded or inaccuracies in the patient details, would require an incident to be entered in the Incident Log such as:

- Attempts to contact participants / volunteers who are now deceased (this is due to not being notified of the status of the participant but is still an IG incident)
- Duplicate participant records
- System inaccessibility
- Database rollbacks and restores

Data management, documentation and curation

Managing, storing and curating data

This **DRAFT** is focussing on data from the IBD BioResource programme.

Currently there are 5 filestores where IBD BioResource data may reside: at AIMES data centre in Liverpool - https://aimes.uk/; at the University of Cambridge High Performance Computing Service (HPC) - https://www.hpc.cam.ac.uk/; in designated SharePoint sites within Microsoft365; on designated areas of the University of Cambridge Clinical School Computing Service network (CSCS) - https://cscs.medschl.cam.ac.uk/; and on paper in a locked cupboard in a locked office on the Cambridge Biomedical Campus. Of these neither CSCS nor the HPC may be used for identifiable data.

These are the main data ingest sources:

1. Consent/Contact details, filled on paper and sent to the BioResource for data entry. It is stored in the recruitment database (CiviCRM).
2. Consent/Contact details/Health & Lifestyle Questionnaire (H&LQ)/Case Report Form (CRF), on paper and sent to the BioResource for data entry. It is scanned using OCR and stored in CiviCRM. All phenotype information is extracted, cleansed and stored in a separate database known as OpenClinica.
3. H&LQ/CRF is also entered by participants using REDCap, an online survey tool and a holding application/data stored via a file storage at AIMES.
4. Consent/Contact details/CRF, participants recruited and registered at the NHS Trusts are registered on local Electronic Health Records. This data is stored at AIMES. All phenotype is then cleansed and stored in OpenClinica. The participant list is reconciled via email with the relevant NHS Trust.
5. Data about samples collected arrives from the laboratory that receives and processes them - the National Biosample Centre at Milton Keynes - and is stored at AIMES.
6. A project is underway to collect genetic data on all participants in the IBD BioResource. Here samples are sent to Thermofisher in the US, and data returned to the HPC.

Data goes through a life cycle: its acquisition is recorded; it is (save the big data in HPC) loaded into audited databases; and curated to make data releases. Those releases are also recorded in detail, and through a data access register at https://bioresource.nihr.ac.uk/studies/?speciality=&studytype=Data%2Bonly&tag= . All data sources are backed up. The AIMES data is snapshotted and stored as encrypted files at AWS, in their London, UK data centre; HPC data is uploaded to the Hinxton, UK instance of the European Genome-Phenome Archive - https://ega-archive.org/ - from where it may be accessed under managed access.

While outside the scope of this **DRAFT**, data for the IBD Registry is also held at AIMES, in an independent tenancy. The IBD Registry routinely uses a Trustworthy Research Environment (TRE) from which data may not be downloaded. The main sources of data ingest are:

1. Patient Reported Outcome Measures (PROMs) received from participants, via REDCap
2. Linked health record data from NHS Digital, based on the record of consenting participants in clinic.

Data for the Wellcome Sanger Centre is held in their own data centre and is processed on their own high performance computing cluster - https://www.sanger.ac.uk/group/information-communications-technology/ For Gut Reaction, a copy of standard file formats generated by the Sanger, and post-QC, are held by the NIHR BioResource at the University of Cambridge High Performance Computing Service - https://www.hpc.cam.ac.uk/

Linkage

Linkage between NIHR BioResource participants and NHS Trusts, is achieved by (securely) reminding recruiting Trusts of the participant identifiers and personal details of their recruits.

Linkage between NIHR BioResource and Wellcome Sanger Centre data is achieved through sharing of identifiers and data, under contract: genetic data is not personal data if it cannot be linked to the

person, which makes sharing data easier where a participant has consented to one party (and is known) and not to the other (and is not known).

Linkage between the IBD Registry and NIHR BioResource, is harder, as the data shared would still be personal data when de-personalised. Linkage is achieved through a method of comparing hashed personal data before data is released: if the hash does not match, it is not the same person. The hash cannot be reversed to re-discover personal details. This privacy-preserving method has been described widely in e.g. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-017-0437-1

Linkage between IBD Registry and Wellcome Sanger Centre, where the participant is also not in the NIHR BioResource, is not possible: the latter has insufficient personal details to create a hash, and the former has no genetic data.

De-personalisation

Having achieved linkage, data is de-personalised to various degrees before any access is approved. Data privacy in Gut Reaction is supported by a set of technologies across the different stages of the pipeline:

1. Confidential data is hosted in a secure data centre run by AIMES.
2. NHS Trust data is hosted on a secured SharePoint environment, with individual access.
3. De-personalisation is supported by pseudo-anonymisation procedures (such as banding, tokenization, replacement of external identifiers, small numbers deletion) along with the use of built-in policies provided by the Privitar software, with workflow managed through Apache Nifi
4. Each research proposal will receive access to a different dataset, water-marked with project-specific identifiers.

Metadata standards and data documentation

Metadata for the primary sources of data is captured in the HDR UK Innovation Gateway - https://www.healthdatagateway.org/ , in the Gut Reaction Hub's own collection - https://web.www.healthdatagateway.org/collection/8070361309216243

Additional documentation is held on the Gut Reaction website - https://gut-reaction.org/data/datasets-available/datasets-available-data-sets-detail/. This includes PDFs of the data capture forms, data catalogues with data profiling. Separately there are Venn diagrams to show the overlap between datasets - https://gut-reaction.org/data/datasets-available/dataset-intersectionality/.

One use of metadata standards of note: the Gut Reaction Cohort Discovery Tool uses data dictionaries to map data into an i2b2 data warehouse - https://www.i2b2.org/ . Data from self-report Health & Lifestyle Questionnaires and clinical Case Report Forms is mapped to SNOMED-CT codes for the purpose of recording diagnoses, procedures and medications. Upper level medication classes are recorded using the Anatomical Therapeutic Chemical (ATC) classification system. Rare Disease abnormalities are coded according to the Human Phenotype Ontology (HPO).

Data preservation strategy and standards

All 3 of the partners described have long-term aspirations for the data they hold:

1. the NIHR BioResource uses data to invite participants from the IBD BioResource to experimental medicine studies

2. the IBD Registry's core business is around clinical evaluation and audit, and changes in patient treatment and outcomes over time
3. the UK IBD Genetics Consortium is building ever larger cohorts of participants to investigate more fine-grained aspects of disease using more subtle genomic techniques.

Therefore, we assume that data we collect will have long-term value.

For the NIHR BioResource, we protect our day-to-day data holdings in three main ways:

- We follow best technical practice in how we handle information:
    - we encrypt data when we have to move it
    - we keep data in secure data centers – both physically secure against intruders, and electronically secure against hackers
    - we keep personal details separate to other forms of information
    - we monitor who can access what.
- We train our staff carefully, so they know what they need to do to keep information safe. We do this to NHS standards, using NHS training materials
- We check these standards are met.

For long-term preservation - and before this **DRAFT** can be submitted - data will be placed in standard formats in managed access repositories. A substantial amount of genetic data is already available (as VCFs and CRAMs) at the EGA - see https://ega-archive.org/dacs/EGAC00001000259

The NIHR BioResource has ethical approval to keep (and therefore allow access to) data for 10 years after the study has finished (to November 2032 in the first instance). Practically, this would involve placing data under the guardianship of Cambridge University Hospitals NHS Foundation Trust (CUH) who are the Data Controller.

Data Records

For this **DRAFT** this section largely describes data managed by the NIHR BioResource. For now, for more details on the datasets and updated data dictionaries, see: https://gut-reaction.org/data/datasets-available/datasets-available-data-sets-detail/

Dataset overview

The following tables summarizes the participant counts across various datasets for Gut Reaction. There are a total of 34,000 participants who have given consent to use their data for this project. The broad datasets along with their participant counts are summarized in **Table 4**.

Dataset	Population Description	Population size	Measured Property	Observation date
NIHR IBD BioResource: Sample Holdings	Number of participants with Blood Samples	20,405	COUNT	12/09/2021
	Number of participants with Plasma isolated	26,084	COUNT	12/09/2021
	Number of participants with Serum isolated	29,431	COUNT	12/09/2021
	Number of participants with DNA isolated	25,940	COUNT	12/09/2021
NIHR IBD BioResource: Contact detail	Participants who have contact details	34,065	COUNT	01/09/2021
NIHR IBD BioResource: Case Report Form	Participants who have clinical data (disease and/or medication data collected at recruitment by a healthcare professional)	33,904	COUNT	01/09/2021
NIHR IBD BioResource: Demographic	Participants who have demographic information (age and gender as a minimum)	33,935	COUNT	01/09/2021
NIHR IBD BioResource: Health and Lifestyle Questionnaire	Participants who returned health and lifestyle questionnaires	22,555	COUNT	01/09/2021
NIHR IBD BioResource: SNP Chip and imputation data	SNP genotyping array chips processed and imputation performed.	11,694	COUNT	01/09/2021

Wellcome Sanger Institute: Whole Exome Sequencing	Whole Exome Sequences - CRAM files and VCF joint files.	6,996	COUNT	01/09/2021
NIHR IBD BioResource: Consent records	Participants who have consent information	34,353	COUNT	01/09/2021

Table 4: counts (as at 9th September 2021) of participants represented within Gut Reaction

Overlap of datasets

Figure 1 shows the relationship between IBD BioResource datasets. Approximately, 7,000 participants link across all core NIHR IBD BioResource datasets for Gut reaction, and the number will grow to over 20,000 as the genetic data builds during 2021 and 2022. This linkage allows rich data analysis for longitudinal studies.

For this **DRAFT**, for more and revised information on how various datasets interlink, see https://gut-reaction.org/data/datasets-available/dataset-intersectionality/

Measurements

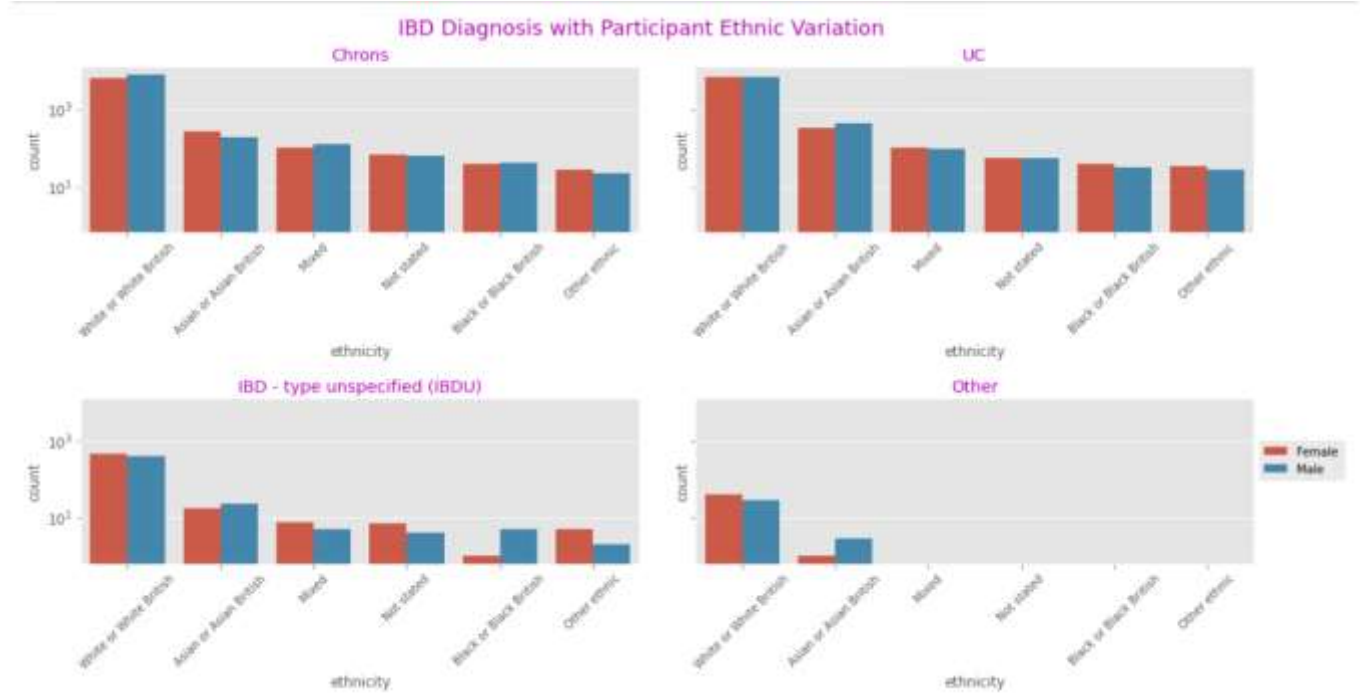
*For this **DRAFT**, this section holds some exploratory data summaries which have not been dignified with Figure or Table headings. As there is usually a limit on submitted figures, and to a lesser extent tables, these will need to be shaped into a few key items.*

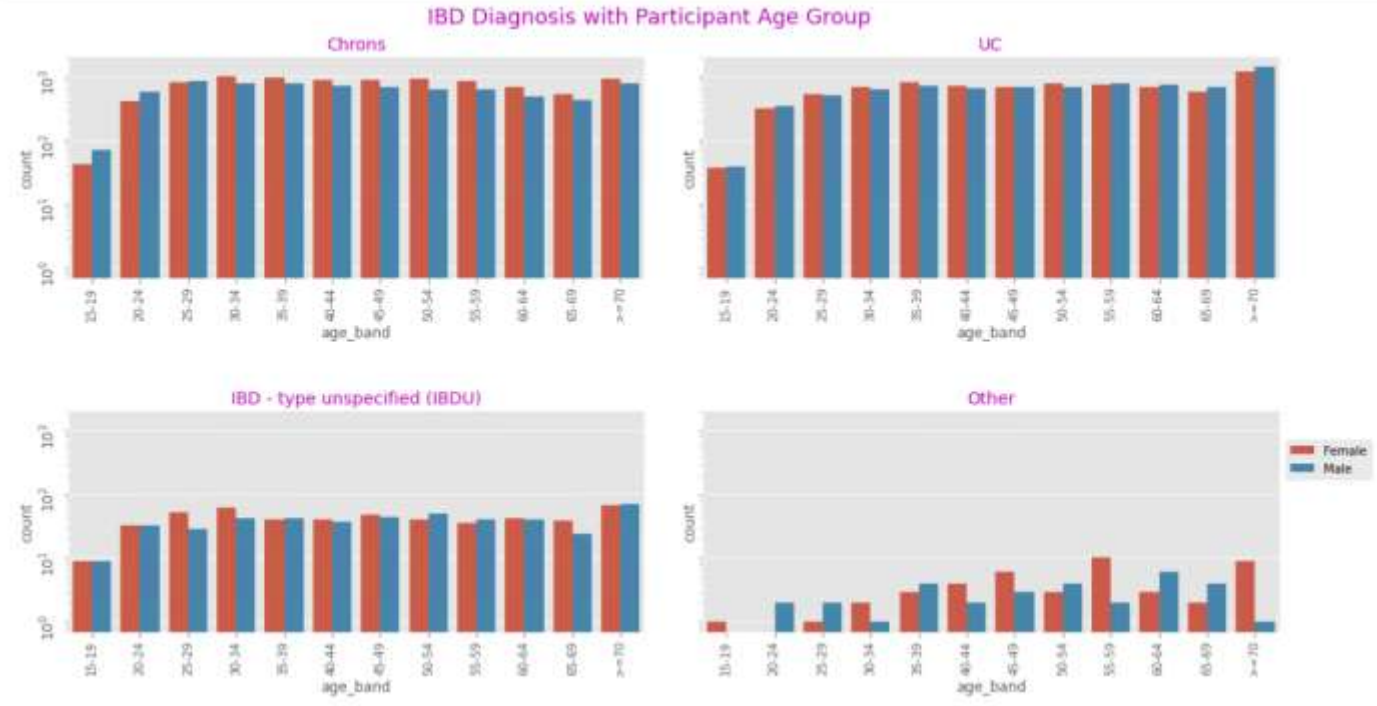
Socio-demographic data

Data captured includes Age, Gender, and Ethnicity.

The following figures show some of this data for various IBD diagnoses. From both these figures we conclude that the IBD Cohort:

- Has a balanced distribution among participant age groups.
- Has a relatively skewed distribution in the ethnic background of participants, however, this is a work in progress and the IBD BioResource and other source institutions are recruiting more participants in other ethnic groups.
- Gender distribution is also well-balanced.

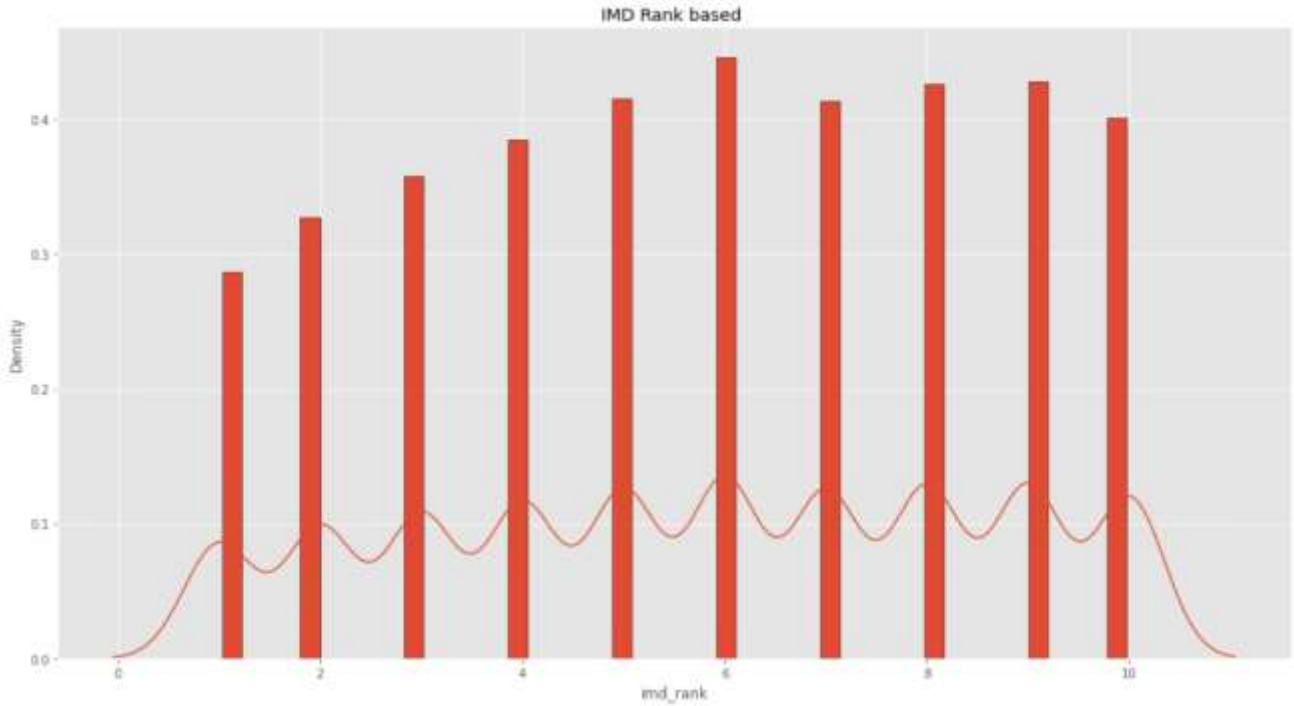


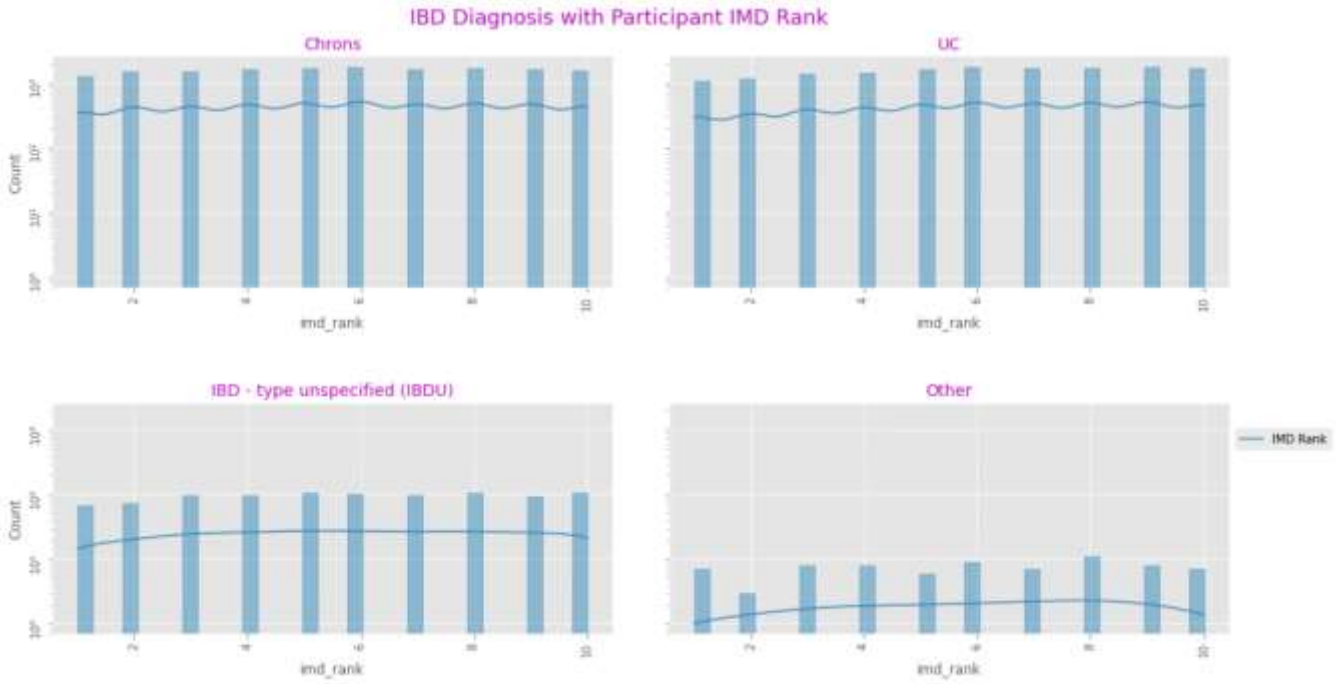


Health and Lifestyle

This data aims to capture the health and lifestyle behaviour of recruited participants. Common fields of interest include information on various co-morbidities for participants (Smoking behaviour, Alcohol consumption, BMI, Dietary habits) and data regarding their quality of life.

While we do not expect our participants to be representative of the population as a whole, from the graph below, we can see that recruited participants have a reasonable distribution on their indices of multiple deprivation (ref), although statistical testing shows that among various IBD Cohorts, the IMD rank would remain an important co-variate.





Measures specific to IBD

The health and lifestyle questionnaire and the clinical case report form contain plenty of IBD Specific conditions diagnosed for participants, and medications prescribed along with adverse reactions to these. Additionally, the clinical case report form also contains details on surgeries for participants. Some of the IBD Specific fields of interests useful to researchers are as follows:

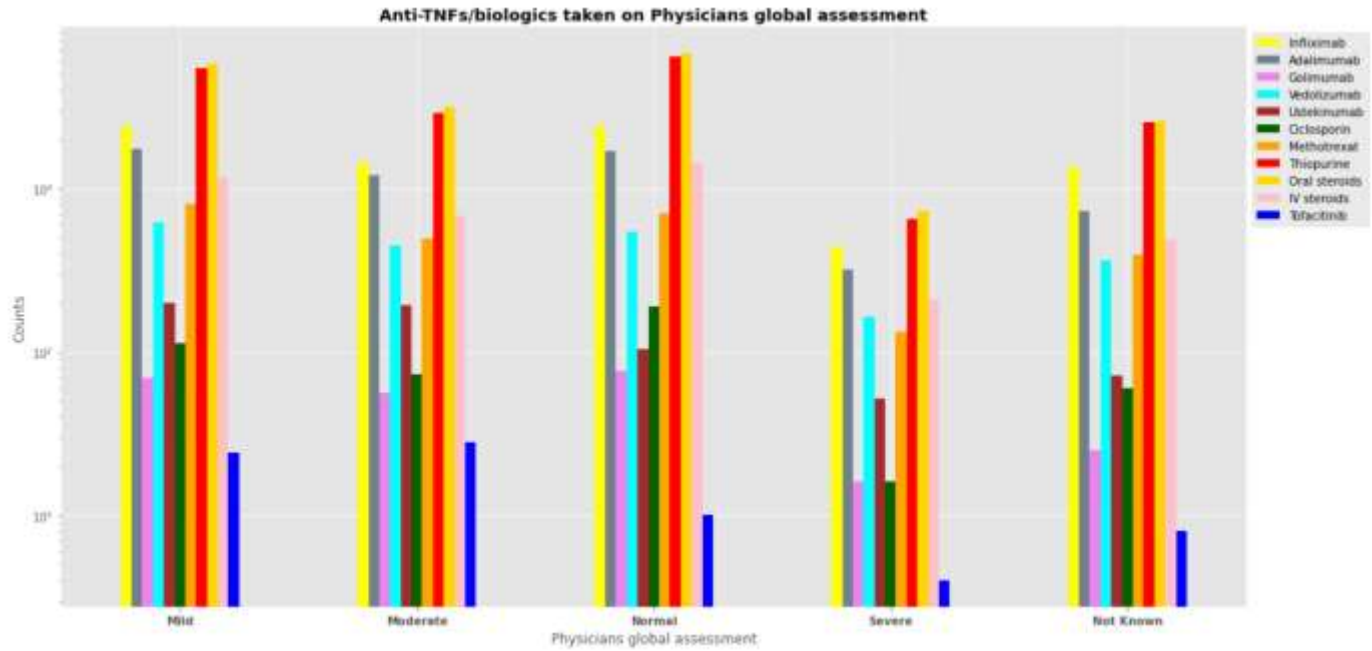
- Date/year of first IBD diagnosis
- Dates of medications prescribed and adverse reactions to these medications.
- IBD inflammatory activity (normal, mild, moderate, severe, unknown) based on the physician's assessment of the IBD.
- Extra-intestinal manifestations & comorbidities.
- Anti-TNFs/biologics taken (Infliximab, Adalimumab, Golimumab, Vedolizumab, Ustekinumab listed separately)
- Amount of time Anti-TNFs/biologics taken for
- Whether currently taking Anti-TNFs/biologics
- Efficacy of medications along with adverse reactions, if any.
- Surgeries and treatments participants go through.

All these fields are captured in the clinical case report forms and the IBD Health and Lifestyle Questionnaires. The sections below give a flavour on some of these fields of interest to clinicians and researchers.

Global Assessment and Drugs

The following table shows participants prescribed anti-TNF biologics based on the physician's global assessment. The datasets also have a rich collection of timestamps when these medications were prescribed and adverse reactions to these medications.

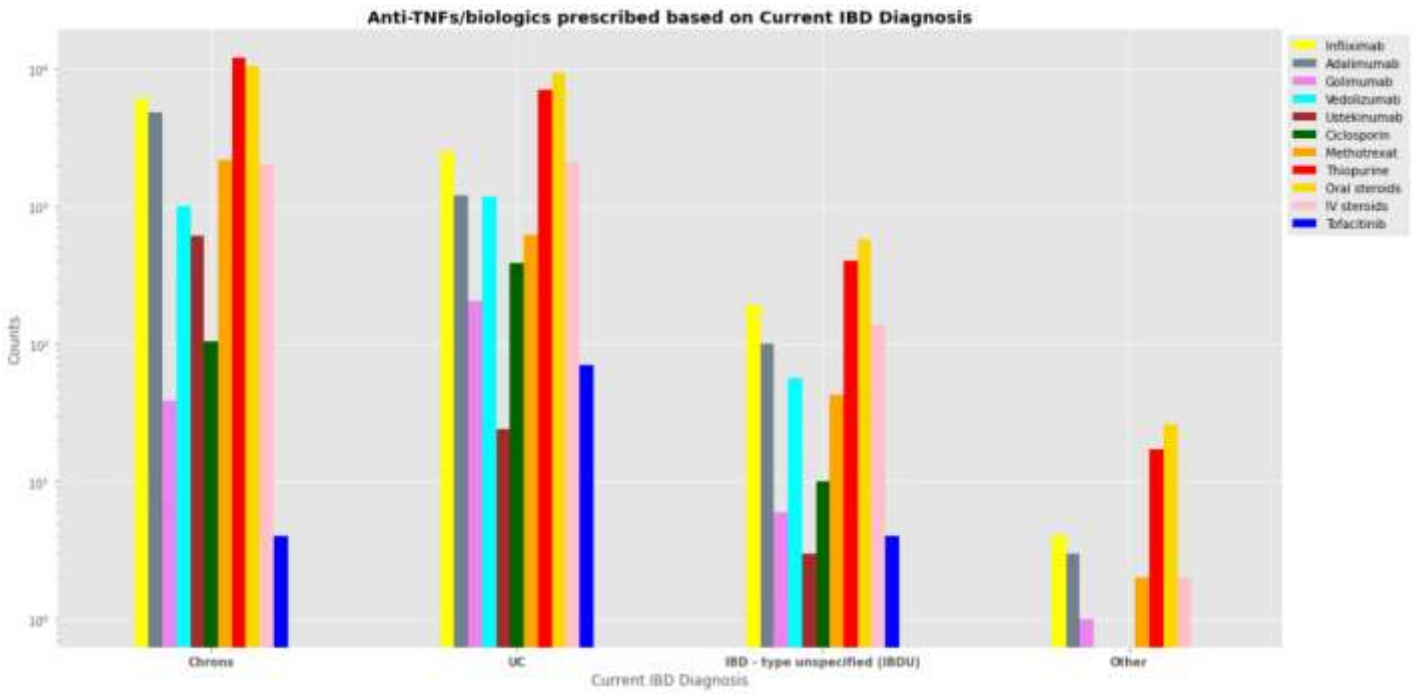
Global assessment	Infliximab	Adalimumab	Golimumab	Vedolizumab	Ustekinumab	Ciclosporin	Methotrexat	Thiopurine	Oral steroids	IV steroids	Tofacitinib
Mild	2420	1750	69	620	201	113	802	5492	5815	1168	24
Moderate	1444	1213	56	450	192	73	495	2900	3158	689	28
Normal	2431	1705	77	541	105	189	707	6430	6713	1439	10
Not Known	1376	731	25	364	72	60	395	2558	2575	486	8
Severe	441	319	16	163	52	16	132	647	730	208	4



Diagnosis and Medications

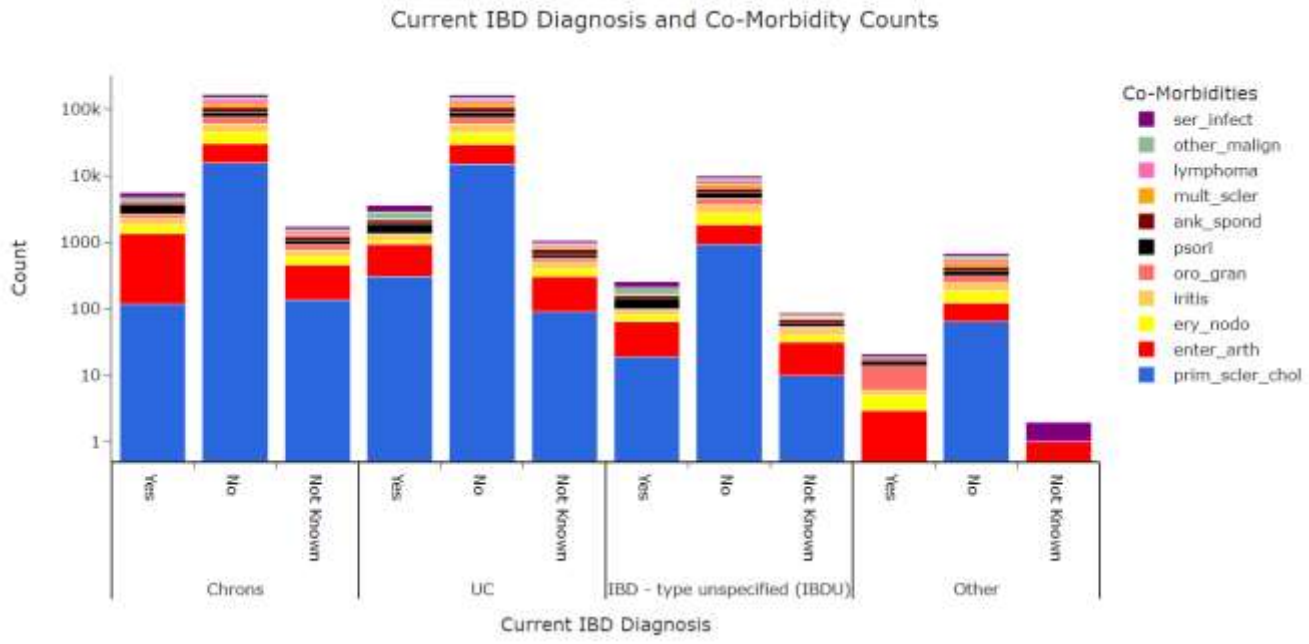
Clinicians can additionally also look at drugs prescribed based on current IBD diagnosis of participants. This information is also captured in these datasets.

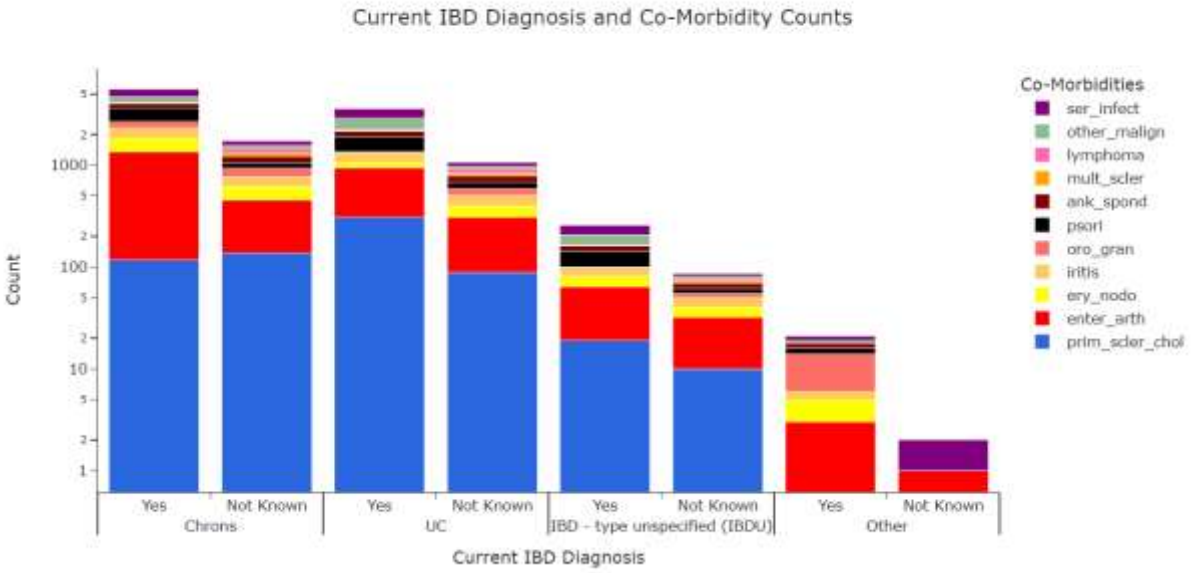
Current IBD Diagnosis	Infliximab	Adalimumab	Golimumab	Vedolizumab	Ustekinumab	Ciclosporin	Methotrexat	Thiopurine	Oral steroids	IV steroids	Tofacitinib
Crohns	6070	4814	39	1011	614	104	2171	11992	10499	2000	4
UC	2511	1205	205	1170	24	385	618	7085	9279	2072	70
IBD - type unspecified (IBDU)	195	101	6	56	3	10	43	399	573	139	4
Other	4	3	1	0	0	0	2	17	26	2	0

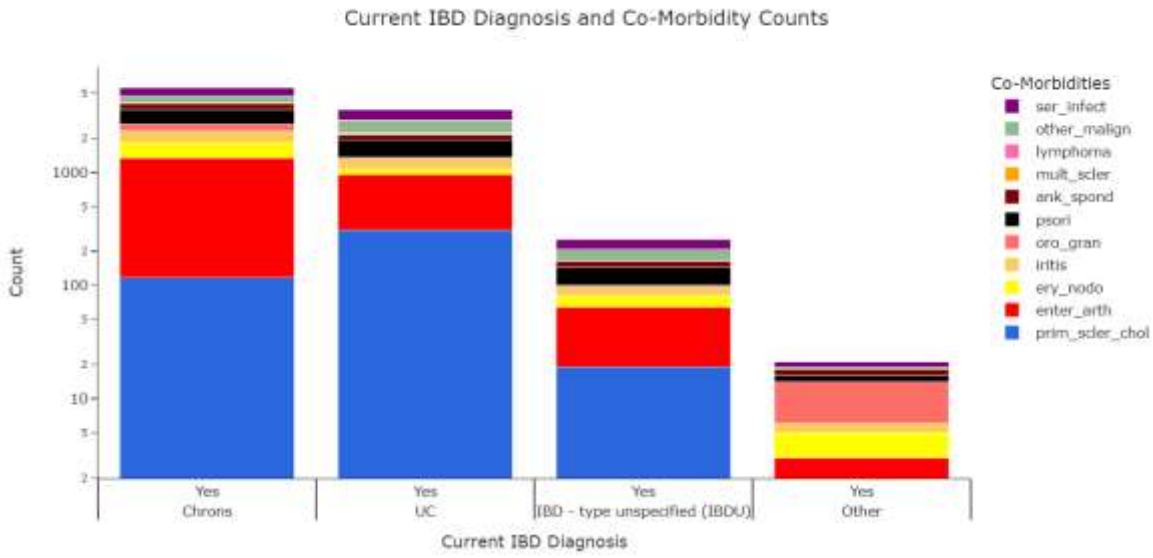


Current IBD Diagnosis and Co-Morbidity

The clinical and the health and lifestyle data also captures a rich array of co-morbidities for participants and captures this information for recruited participants. We can see from the graphs below that most participants have one or more co-morbidity and most IBD Cohorts have Primary Sclerosing Cholangitis and Enteropathic arthritis.



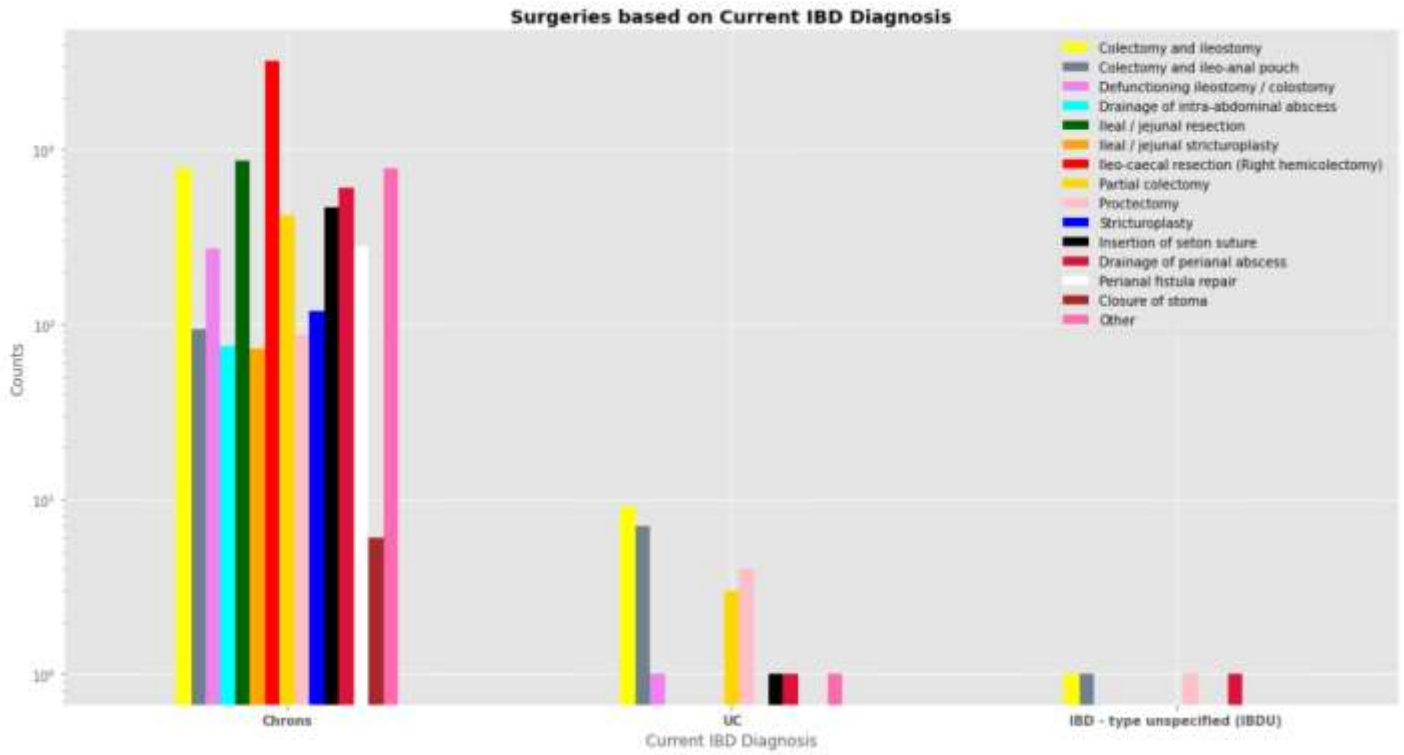

Current IBD Diagnosis and Co-Morbidity Counts
Co-Morbidities
ser_infect
other_malign
lymphoma
mult_scler
ank_spond
psori
oro_gran
iritis
ery_nodo
enter_arth
prim_scler_chol
Count
5
2
1000
5
2
100
5
2
10
5
2
1
Yes
Not Known
Chrons
Yes
Not Known
UC
Yes
Not Known
IBD - type unspecified (IBDU)
Yes
Not Known
Other
Current IBD Diagnosis


Current IBD Diagnosis and Co-Morbidity Counts
Co-Morbidities
ser_infect
other_malign
lymphoma
mult_scler
ank_spond
psori
oro_gran
iritis
ery_nodo
enter_arth
prim_scler_chol
Count
5
2
1000
5
2
100
5
2
10
5
2
Yes
Chrons
Yes
UC
Yes
IBD - type unspecified (IBDU)
Yes
Other
Current IBD Diagnosis

It can be seen that some people have additional co-morbidities alongside their IBD diagnosis and this information can be extremely useful for clinical research.

Current IBD Diagnosis and Surgeries

There is also information on surgeries based on the participant diagnosis. The Clinical data captures these different operations.



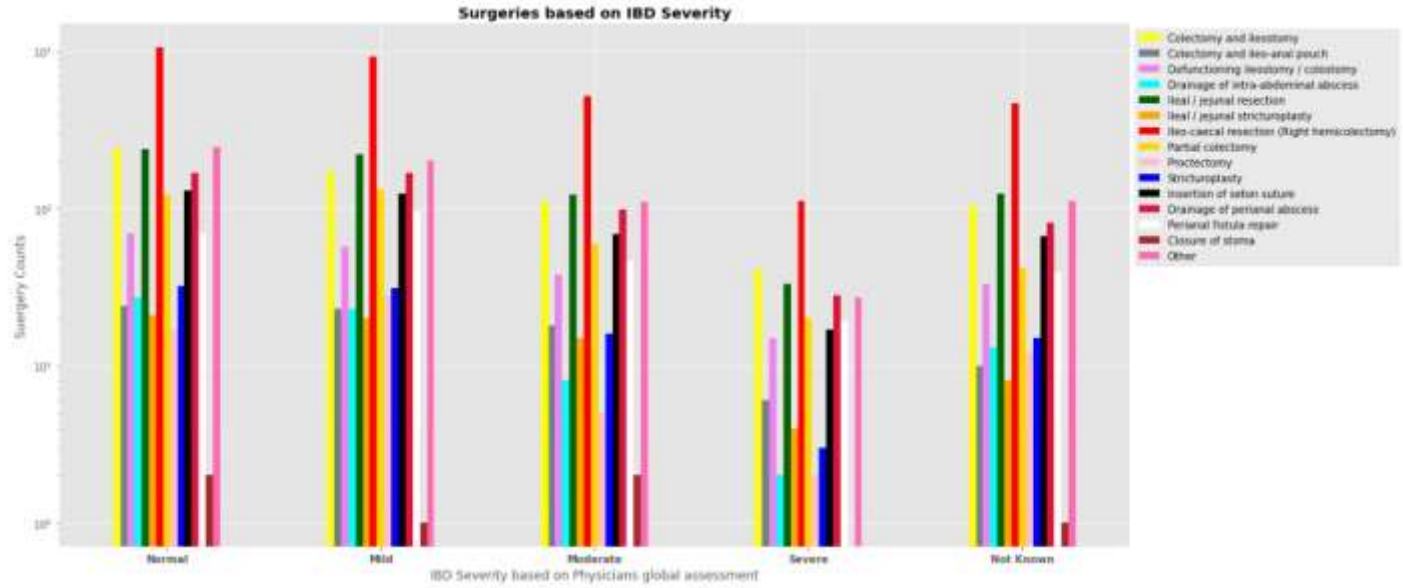
Criteria	Current IBD Diagnosis			
Operations == 1	**Crohns**	**UC**	**IBD - type unspecified (IBDU)**	**Other**
Above	709	7	1	0
Below	15866	15874	974	76

Using Chi-square test, we get the following results to check for statistical significance. Statistically, there is a relation among the surgeries prescribed for the current IBD diagnosis.

Degrees of Freedom	Chi-Square Statistic	P-value
3	715.7878215411453	7.91748281708055e-155

Global assessment and Surgeries

Surgeries prescribed are also available based on the physician's global assessment of IBD.



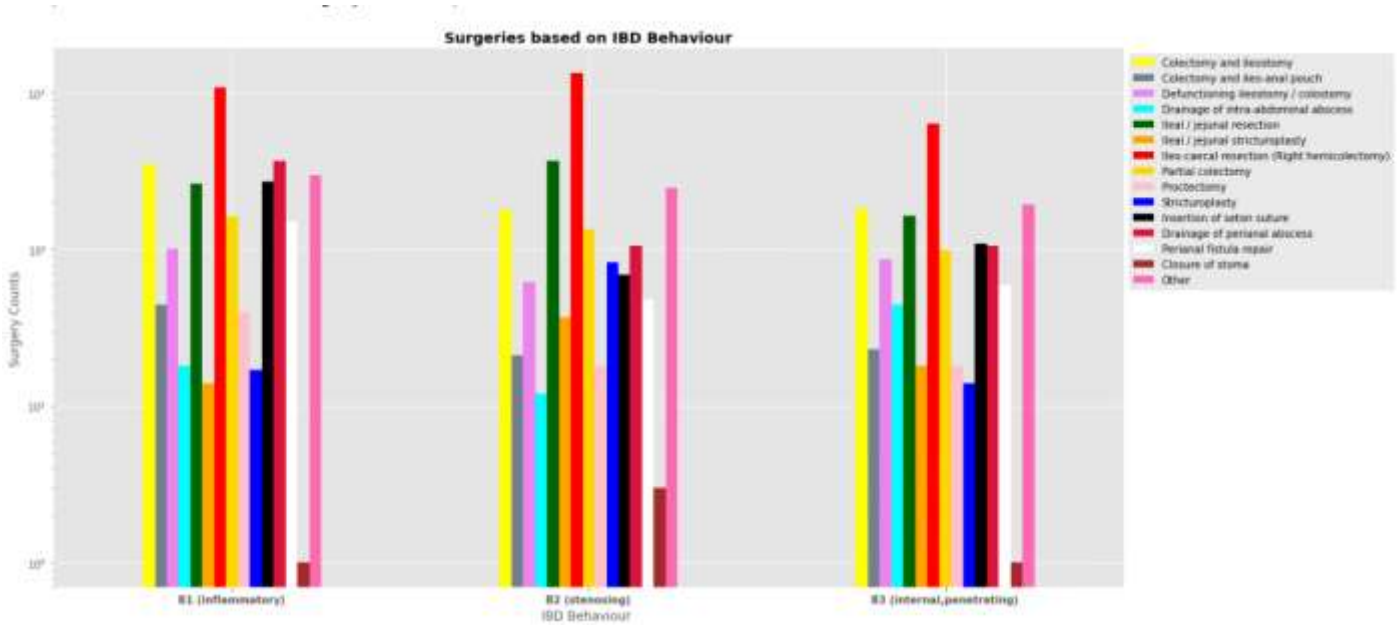
Criteria	Global Assessment				
Operations > 1	**Mild**	**Moderate**	**Normal**	**Not Known**	**Severe**
Above	178	76	198	92	33
Below	9262	4457	11533	4134	952

Using Chi-square test, we get the following results to check for statistical significance.

Degrees of Freedom	Chi-Square Statistic	P-value
4	17.019885022526605	0.0019158293296061212

Surgeries and IBD presentation

Surgeries undertaken are also available based on the nature of the IBD.



Criteria	IBD presentation ("Behaviour" part of Montreal classification)		
Operations > 1	**B1 (Inflammatory)**	**B2 (Stenosing)**	**B3 (Internal Penetrating)**
Above	266	197	210
Below	9512	3561	1680

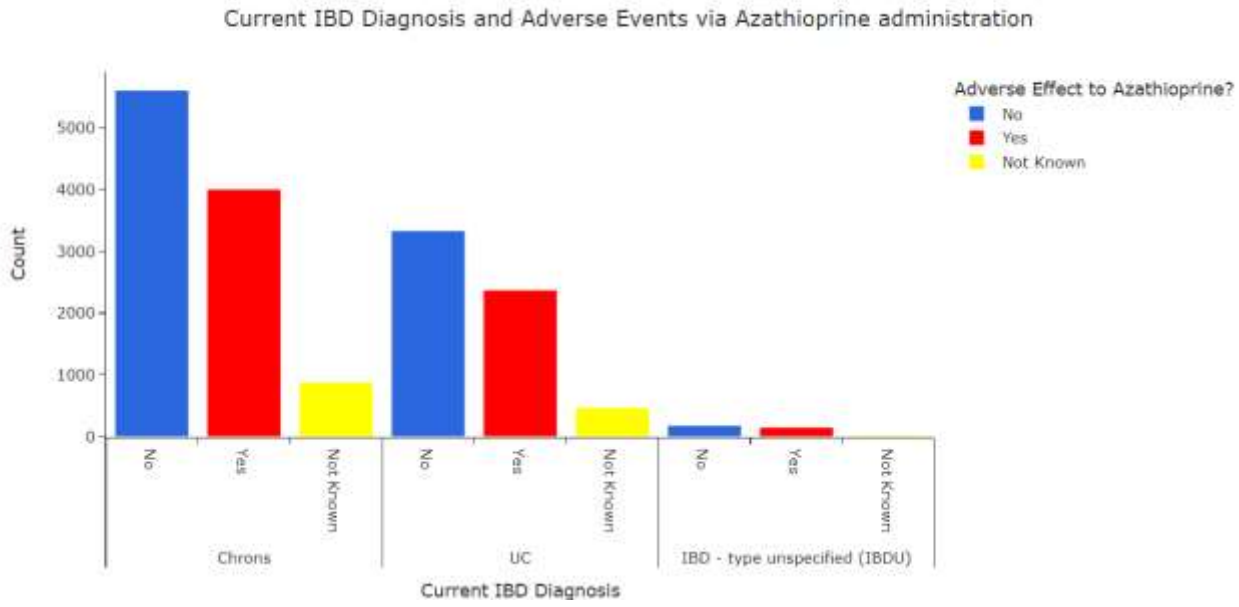
Using Chi-square test, we get the following results to check for statistical significance.

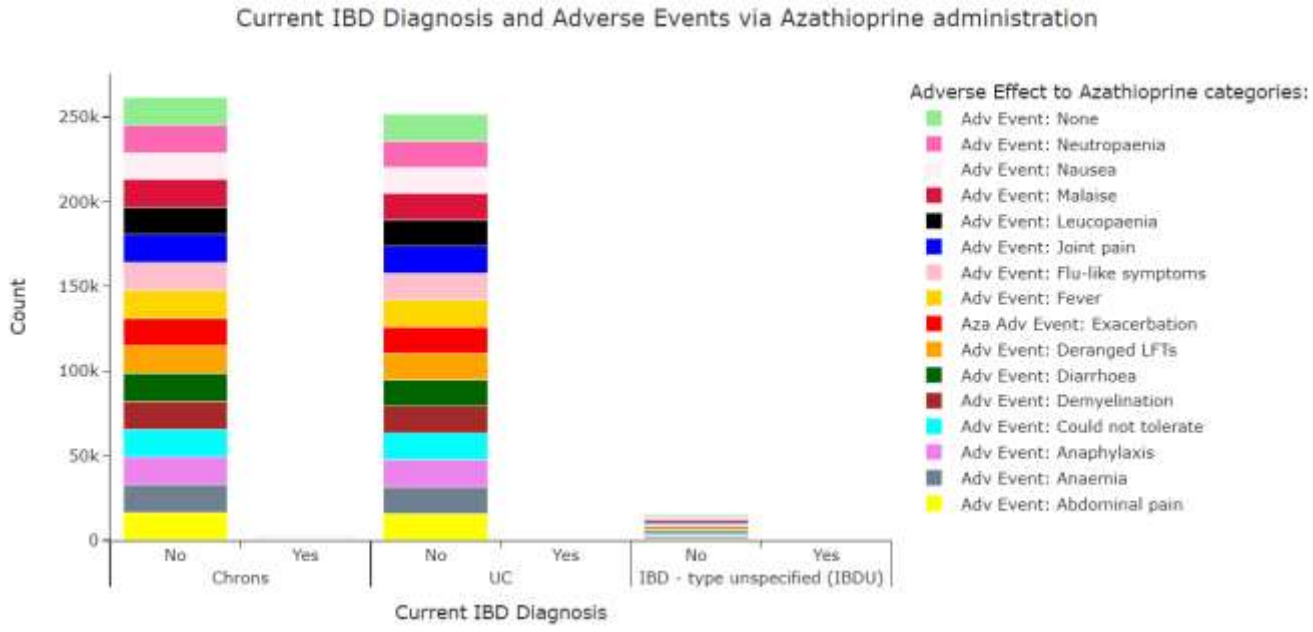
Degrees of Freedom	Chi-Square Statistic	P-value
2	276.46	9.265196490944545e-61

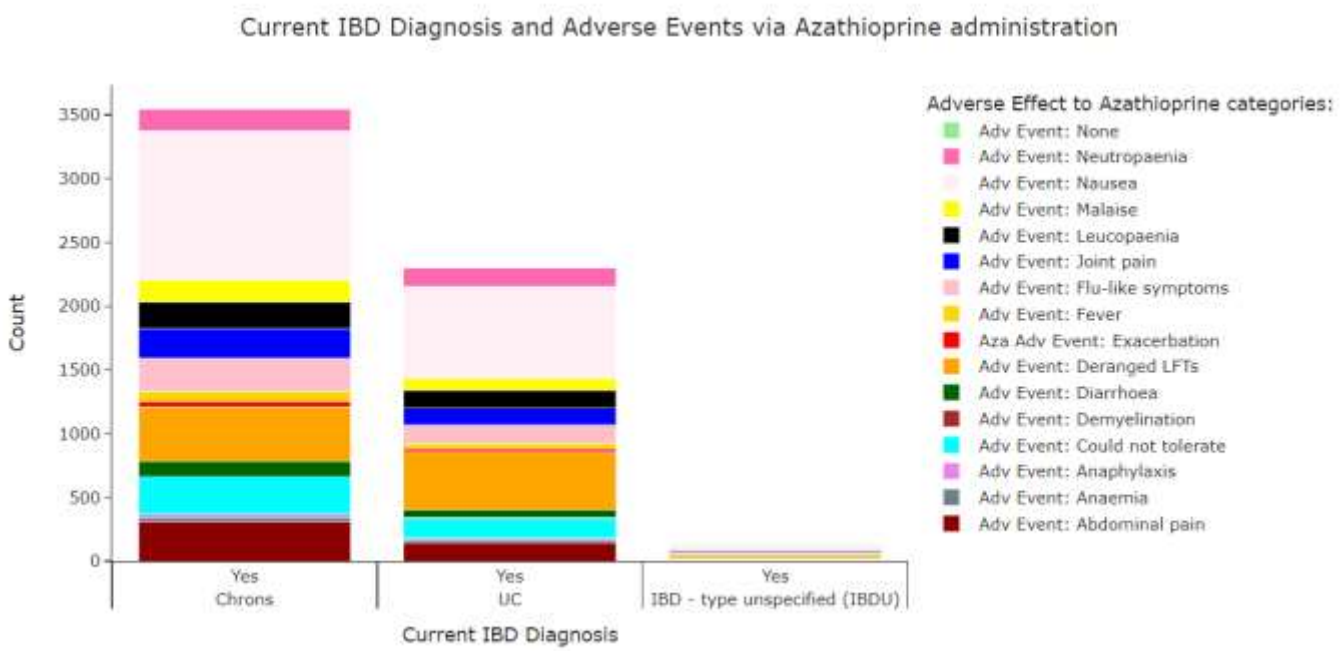
This demonstrates, if any demonstration is required, that IBD presentation has an impact on the number of surgeries participants go through. This attribute could be a good predictor for machine learning based predictive analytics. We will also try to verify these classifications from the data.

Medication Adverse Events

The clinical case report form also captures adverse events to various IBD medications prescribed to participants, logging their symptoms like nausea, abdominal pain etc.


Current IBD Diagnosis and Adverse Events via Azathioprine administration
Adverse Effect to Azathioprine?
No
Yes
Not Known
Count
5000
4000
3000
2000
1000
0
No
Yes
Not Known
Chrons
UC
IBD - type unspecified (IBDU)
Current IBD Diagnosis


Current IBD Diagnosis and Adverse Events via Azathioprine administration
Adverse Effect to Azathioprine categories:
Adv Event: None
Adv Event: Neutropaenia
Adv Event: Nausea
Adv Event: Malaise
Adv Event: Leucopaenia
Adv Event: Joint pain
Adv Event: Flu-like symptoms
Adv Event: Fever
Aza Adv Event: Exacerbation
Adv Event: Deranged LFTs
Adv Event: Diarrhoea
Adv Event: Demyelination
Adv Event: Could not tolerate
Adv Event: Anaphylaxis
Adv Event: Anaemia
Adv Event: Abdominal pain
Count
250k
200k
150k
100k
50k
0
No
Yes
Chrons
UC
IBD - type unspecified (IBDU)
Current IBD Diagnosis



Genetic Datasets

The NIHR IBD BioResource extracts DNA from blood and saliva samples taken at recruitment, and measures a panel of Single Nucleotide Polymorphisms (SNPs) on each DNA sample, using a commodity SNP genotyping array from e.g., Illumina or Affymetrix (now Thermofisher). This is used to pre-screen or match participants when inviting them to take part in experimental medicine studies.

The Annotation file for SNP Chip – modelled on that used by the UK Biobank (ref) is available here: https://bioresource.nihr.ac.uk/media/1o3c4m42/axiom_ukbbv2_1-na36-r3-a4-annot-csv.zip

The imputation method for SNP chips is described here: http://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/impute_ukb_v1.pdf

There is also a substantial overlap between the participants in the NIHR IBD BioResource and the long-running IBD UK Genetics Consortium (IBDGC). The Wellcome Sanger Institute performs the sequencing work for the IBDGC. The NIHR BioResource provides DNA samples to this initiative. For those samples, and where there is an additional overlap, e.g., because a participant has been seen before and therefore no new sample is required, this data is being provided to the Gut Reaction Hub by the Wellcome Sanger Institute.

The following table summarises the genetic data available for the Gut reaction Hub.

Dataset	Dataset descriptor	Population Size	Observation date
Wellcome Sanger Institute: Whole Exome Sequencing	Whole Exome Sequences - CRAM files	6,996	01/09/2021
	Whole Exome Sequences - VCF joint file	6,996	01/09/2021
	SNP genotyping array chips processed	11,787	01/09/2021

NIHR IBD BioResource: SNP chip data and Imputation			
	Approximate number of variants on a SNP chip	780,000	01/09/2021

Technical Validation

Recent publications using this data are available here:

1. Impact of NOD2 Genetic Variants on the Gut Mycobiota in Crohn's Disease Patients in Remission and in Individuals Without Gastrointestinal Inflammation | Journal of Crohn's and Colitis | Oxford Academic (oup.com)

2. Thiopurine monotherapy is effective in ulcerative colitis but significantly less so in Crohn's disease: long-term outcomes for 11 928 patients in the UK inflammatory bowel disease bioresource | Gut (bmj.com)

3. IBD BioResource: an open-access platform of 25 000 patients to accelerate research in Crohn's and Colitis | Gut (bmj.com)

4. Impact of NOD2 Variants on Fecal Microbiota in Crohn's Disease and Controls Without Gastrointestinal Disease | Inflammatory Bowel Diseases | Oxford Academic (oup.com)

Usage Notes

Information on how researchers and academics can apply and use the data for research is available in the links below:

- Industry BioResource Usage Costs (nihr.ac.uk)
- BioResource usage costs for academic and clinical researchers (nihr.ac.uk)

Datasets are published quarterly with a time lag of 1-2 months and the averaging processing time of most datasets is between 2 to 6 months. The format in which data is available varies based on data types. Imaging data is available via XNAT servers. Clinicians also have access to a cohort discovery tool in the Trust worthy research environment at AIMES, which can be used to get specific IBD patient cohorts. All datasets follow the de-personalisation policy applied via the Privitar software. Some de-personalise data released to researchers in csvs, excel files, if it is platform-independent. However, access to any data acquired via NHS Digital is subject to strict restrictions governing where data may be accessed and from which locale - access in a Trustworthy Research Environment (TRE) at AIMES secure data centre (https://aimes.uk/).

Code Availability

Code will be available from the University of Cambridge's Gitlab repository – which will need to be made public before this **DRAFT** is submitted.

Acknowledgements

The NIHR BioResource acknowledges all participants involved in clinical research and studies.

NIHR BioResource. Acknowledgement text: *"We thank NIHR BioResource volunteers for their participation, and gratefully acknowledge NIHR BioResource centres, NHS Trusts and staff for their contribution. We thank the National Institute for Health Research, NHS Blood and Transplant, and Health Data Research UK as part of the Digital Innovation Hub Programme. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care."*

Additional funder acknowledgements - grant numbers needed to progress beyond **DRAFT** - include:

- NIHR
- MRC
- Crohn's & Colitis UK

Additional acknowledgements will be made to all partner organisations who attend the Gut Reaction Programme Board, but are not represented in the final authorship list – e.g.:

- Wellcome Sanger Institute
- All NHS Trusts involved in Gut Reaction.

Author contributions

This **DRAFT** has the following authors:

1. Yuchun Ding – responsible for many of the analyses presented
2. Aneeq Rehman – responsible for analysis, the Venn diagram, and much text especially relating to dictionaries and metadata
3. Alvaro Ulrich – responsible for most of the text on data management
4. Neil Walker - editor

Competing interests

The authors declare they have no competing interests.

References

For this **DRAFT** – to be done: added where missing, and taken out of inline text.